

# Incorporation of Speech Duration Information in Score Fusion of Speaker Recognition Systems

Ali Khodabakhsh, Seyyed Saeed Sarfjoo,

Osman Soyyigit, Cenk Demiroğlu

Electrical and Computer Engineering Department

Özyeğin University, 34794, Cekmekoy, Istanbul, Turkey

Email: {ali.khodabakhsh, saeed.sarfjoo, osman.soyyigit}@ozu.edu.tr,  
cenk.demiroglu@ozyegin.edu.tr

Umut Uludag

TUBITAK BILGEM, 41470, Gebze, Kocaeli, Turkey

Email: umut.uludag@tubitak.gov.tr

**Abstract**—In recent years identity-vector (i-vector) based speaker verification (SV) systems have become very successful. Nevertheless, environmental noise and speech duration variability still have a significant effect on degrading the performance of these systems. In many real-life applications, duration of recordings are very short; as a result, extracted i-vectors cannot reliably represent the attributes of the speaker. Here, we investigate the effect of speech duration on the performance of three state-of-the-art speaker recognition systems. In addition, using a variety of available score fusion methods, we investigate the effect of score fusion for those speaker verification techniques to benefit from the performance difference of different methods under different enrollment and test speech duration conditions. This technique performed significantly better than the baseline score fusion methods.

**Index Terms**—speaker recognition, i-vectors, score fusion, short-duration

## I. INTRODUCTION

Over recent years, following the success of the identity vector (i-vector) based speaker verification (SV) methods [1], these systems have made significant progress [2]. As speaker recognition technology reaches its maturity, real-life applications impose a drastic limitation for the systems in terms of environmental noise and speech duration variability during authentication. The problem of duration variability is known to be one of importance for practical speaker recognition applications, and has also been addressed to a certain extent in the literature in the context of i-vector based speaker recognition systems [3]. Furthermore, in the biannual speaker recognition evaluation (SRE) challenge held by National Institute of Standards and Technology (NIST), in year 2014, NIST coordinated a special i-vector challenge [4], where the duration variability was one of the dominant challenges.

Most of the studies on i-vector based speaker recognition focus on recognition problems, where i-vectors are extracted from speech recordings of sufficient length. Therefore, the majority of modeling techniques simply assume that the extracted i-vectors give a reliable estimation of the attributes of the speaker. However, since the duration of recordings can be very short (on the order of less than 5 seconds) in many real-life applications, such as speaker identification tasks in broadcast data, this assumption fails to hold.

Only recently, a number of solutions have been proposed addressing the problem of duration variability. For example, in [3], [5], and [6], authors do not treat i-vectors as point

estimates of the hidden variables in the eigenvoice model, but rather as random vectors. In this slightly different perspective, the i-vectors appear as posterior distributions, parameterized by the posterior mean and the posterior covariance matrix. In [7] Borgstrom and McCree proposed a framework of super vector bayesian speaker comparison (SV-BSC), which keeps account of the observation noise throughout modeling and scoring. In [3] with expanding the work in [7] Garcia-Romero and McCree reformulate SV-BSC and reach to SV-PLDA that facilitates the use of practical techniques, such as length normalization, and multi-cut enrollment averaging. In [8] Vesnicer et al. address the problem of duration variability through weighted statistics and demonstrate how established feature transformation techniques regularly used in the area of speaker recognition, such as PCA or WCCN, can be modified to take the duration into account.

In [9], Rastoceanu and Lazar make a comparison of different features and methods for score fusion for an independent speaker verification application. In this paper, scores obtained with several types of features were fused with combination methods such as mean, max, min, weighted sum, and classification methods such as: support vector machines (SVM), linear discriminant analysis (LDA). Based on the result of this paper, fusion methods outperformed the baseline GMM-UBM method. In [10], for fusion of different classifiers in speaker verification systems, Hautamaki et al. use classifier ensemble selection, which can be seen as sparse regularization applied to logistic regression. However, none of the mentioned studies take duration variability into account for score fusion.

In this paper we investigate the effect of speech duration on the performance of three speaker recognition systems representing the current state-of-the-art: Gaussian Mixture Model-Universal Background Model (GMM-UBM) [11], Total Variability Space (TVS) [1], [12] and Probabilistic Linear Discriminant Analysis (PLDA) [13] scoring in TVS. Furthermore, using a wide range of available score fusion methods, we investigate the effect of score fusion of these speaker verification techniques, to benefit from the performance difference of different methods under different enrollment and test speech duration conditions.

This paper is organized as follows. A background on speaker recognition, and description of speaker recognition systems and score fusion methods used in this study are given

in Section II. Experimental setup and results are presented and discussed in Sections III and IV, respectively. Finally, conclusion and future works are given in Section V.

## II. OVERVIEW OF SPEAKER RECOGNITION SYSTEMS

### A. Feature extraction

Feature extraction as transformation of the speech signal to a set of feature vectors representing the desired attribute, can be done using several different features. In this study, similar to most studies involving speech and speaker recognition, the Mel-Frequency Cepstral Coefficients (MFCCs) are used due to their better performance compared to other features [14].

Sometimes to improve the robustness of features to channel differences, feature normalization methods such as cepstral mean subtraction [15], cepstral mean and variance normalization, and feature warping are used. In this study we used feature warping [16].

To increase the quality and effectiveness of modeling, non-speech frames need to be discarded prior to modeling. Here, voice activity detection (VAD) is done using bi-gaussian modeling of speech frames on log energy distribution of the input frames.

### B. Speaker Modelling

1) *GMM-UBM*: GMMs are typically used to represent the acoustic feature space in speaker recognition systems [17].

In this method, first, using expectation-maximization (EM) algorithm [18], a GMM called universal background model (UBM) is trained from multiple sessions from multiple speakers. For the enrollment step, given an utterance, the speaker model is adapted from UBM using maximum posterior adaptation (MAP) [19]. Typically, only the mean is adapted.

2) *Total variability space (TVS)*: The supervector of mean vectors in UBM is of a very high dimensionality, and the number of parameters to adapt is very high. Assuming speech consists of a speaker factor and an additive channel factor, speech model can be formulated as

$$M_s = S + C, \quad (1)$$

where  $M_s$  is speaker and channel dependent supervector,  $S$  is speaker dependent supervector and  $C$  is channel dependent supervector. Eq. 1 can be rewritten as [17]

$$M_s = M_0 + Vy + Ux + Dz, \quad (2)$$

where  $M_0$  is speaker and channel independent mean supervector,  $V$  is a low-rank eigenvoice matrix representing speaker space,  $U$  is a low-rank eigenchannel matrix representing channel space, and  $D$  is diagonal matrix which is modeling the Gaussian noise, and  $x$ ,  $y$ , and  $z$  are random vectors with a standard normal prior [17]. These vectors can be jointly computed using the joint factor analysis (JFA) approach.

In [1] Dehak et al. have shown that even though JFA is successful in increasing the performance of the recognition systems, there remains speaker variability in the channel factor. They proposed a method combining both factors in a single matrix known as T matrix.

$$M_s = M_0 + Tw \quad (3)$$

where  $M_0$  is speaker and channel independent mean supervector,  $T$  is a rectangular matrix of low rank, and  $w$  is the identity vector (i-vector) and is a random variable with an standard posterior distribution. In this approach, the speaker and channel factors are combined into a single vector  $w$  in a lower dimensional space, postponing the speaker and channel factor separation task.

### C. Scoring

For GMM-UBM system, given a set of feature vectors  $X$  and a speaker model  $M_{hyp}$ , the similarity score is computed as

$$score = \log p(X|M_{hyp}) - \log p(X|M_0) \quad (4)$$

For TVS system, an i-vector is extracted from the test utterance, and compared to the i-vector extracted from the enrollment utterance using similarity measures. In this study, similarity comparison is done using cosine distance scoring, and probabilistic linear discriminant analysis (PLDA) [6]. LDA and length normalization were used prior to PLDA in this study [12].

One of the known methods that improve the accuracy of speaker recognition systems, and emphasized by the speaker recognition community, is score fusion of multiple subsystems. Score fusion takes advantage of the fact that different systems make different mistakes, and by combining their output scores, the overall system can reduce the dependence of output decisions on the mistakes of a particular system.

In this study, we focus on simple mean, logistic regression (LR) and neural networks (NN) methods.

## III. EXPERIMENT SETUP

All systems in the experiments were trained with 19 dimensional MFCC features plus log-energy coefficient along with their delta and delta-delta parameters. 25 msec window with 10 msec window shift is used for feature extraction. Static log-energy feature is excluded, making the final dimensionality of features 59. Feature warping is done on each 300 frames (3sec windows). Bi-gaussian VAD is done using same windowing parameters as feature extraction. Details of training, development and test experiments are shown below.

### A. Training

For training the speaker recognition system, VoxForge online corpus [20] which is a user generated corpus is used. This decision was motivated due to huge number of speakers and short-duration of the utterances in this corpus. The corpus consists of many English dialects from native and non-native speakers. Due to a big unbalance between the dialects and genders, we decided to limit the study to only male speakers with North American English dialect. A summary of the training data is given in Table I.

This user generated corpus consists of two parts, registered speakers and anonymous speakers. Speech from registered users, were labeled based on their usernames and used for the supervised modeling steps (LDA and PLDA). For models that did not require labels, the whole data was used. Furthermore,

TABLE I  
DATABASES, NUMBER OF SPEAKERS, AND NUMBER OF SESSIONS THAT  
WERE USED FOR TRAINING, DEVELOPMENT, AND TEST.

	Training		Devel	Test
	Labeled	Unlabeled		
Corpus	VoxForge		WSJ0	WSJ1
Sessions	19009	9637	822	2625
Speakers	521	-	66	152

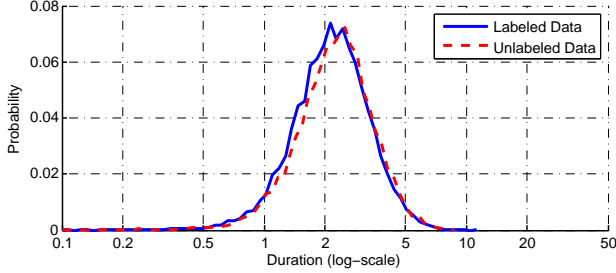


Fig. 1. Normalized histogram of distribution of training data (log-scale).

the maximum number of sessions per speaker was limited. Distribution of these sets are shown in Fig. 1.

A UBM consisting of 1024 gaussians was trained on all the available data. The same data was used for training of the T matrix with a rank of 500. The ranks of LDA and PLDA models were set to 150 and 75 respectively, according to the best performance achieved on the training data.

#### B. Development and test

To minimize the effect of channel difference and to keep the focus on duration variability, Wall Street Journal (WSJ0, and WSJ1) corpus were chosen as development and test sets respectively. To be able to analyze the effect of speech duration, a wide range of enrollment and test speech duration were targeted (0.1, 0.2, 0.5, 1, 2, 5, 10, 20, and 50 seconds). The motivation behind the aforementioned duration setup was that typically segment duration distribution follows a log normal distribution.

To generate this data, speech data of each speaker from WSJ corpus was first concatenated, then split to sessions containing 50 seconds of speech. For each speaker, one session is chosen for enrollment and maximum of 50 of the remaining sessions were used for test/development. As we want to investigate the effect of different enrollment and test durations, we randomly generate 8 shorter sessions from each session with approximately 0.1 to 20 seconds of human speech. For applying a feature warping, minimum recording duration of each part should be more than 3 seconds. By doing this, 9 set of enrollment scenarios and 9 set of test/development scenarios are created.

Details about number of speakers and number of sessions for development and tests data can be found in Table I.

For the sake of simplicity, only T-norm score normalization is done on the output of the subsystems [21]. All identification tests are done under closed-set conditions. Logistic regression (LR) score fusion is done using BOSARIS toolkit [22]. For neural network (NN) score fusion, a feed-forward network consisting 4 nodes in the hidden layer was used.

## IV. RESULTS AND DISCUSSION

Two sets of experiments were done on the test data. In the first set of experiments, all 81 sets of data were combined and accuracy of each subsystem as well as the results for score fusion using mean, LR, and NN were calculated. Score fusion models were trained using the development data. In the second experiment, LR and NN score fusion models were trained on each enrollment and development duration condition separately, creating 81 model for each system. Accordingly, the tests were done using the models trained with the corresponding enrollment and test duration, and accuracies were reported on the whole data. These results are shown in Table II. It is important to note that the high error rates in this table are caused by the very short tests durations (0.1, 0.2, 0.5, and 1 seconds). In fact, experiments with very short durations on at least one side of enrollment or test cover 69% of the total experiments.

One might expect that training a score fusion model for each duration condition will reduce the performance of the fused scores by limiting the amount of training data. However, as it is seen on the Table II, the best performance achieved for speaker verification system is achieved with duration-based system, reducing the error by 1.75%.

To gain insight on the reason for this behavior, a third set of experiments are done on each duration condition, and results were reported independently. In Fig. 2a the error rates of the GMM method is shown. As expected, as the duration of test and enrollment increases, the error rate becomes lower. Due to space saving, only the results for speaker identification experiments are visualized.

To observe the effectiveness of TVS-cosine method, the relative error reduction (RER) rate of this system compared to the base GMM system is shown in Fig. 2b. It can be seen that the biggest relative increase of accuracy occurred around 2 second enrollment and test, which correspond to the training distribution shown in Fig. 1. On the other hand, there is a stable increase of accuracy where the length of enrollment and test is very short, which corresponds to better modeling of TVS-cosine for these situations due to reduction of number of parameters for adaptation as mentioned before. Another interesting observation made is that the TVS-cosine fails to increase accuracy in most cases when the length of enrollment is longer than 20 seconds.

The relative error reduction of TVS-PLDA compared to TVS-cosine is visualized in Fig. 2c. Here a similar effect to previous case can be observed. These results show the dependence of the accuracy gain of PLDA to the duration of enrollment and test, as well as the distribution of training data.

To see how the duration-based LR takes advantage of the dependency of performance of different systems to durations, the relative error reduction of duration-based LR compared to LR is visualized in Fig. 2d. The normalized weights of 0.17, 0.37, and 0.46 were assigned to GMM, TVS-cosine, and TVS-PLDA respectively by the LR model.

TABLE II

SPEAKER IDENTIFICATION ERRORS AND SPEAKER VERIFICATION EQUAL ERROR RATES (EER) ARE SHOWN FOR ALL SUBSYSTEMS AND SCORE FUSION METHODS. RESULTS WITH DURATION-BASED SCORE FUSION ARE ALSO REPORTED WHERE APPLICABLE. SYSTEMS WITH BEST PERFORMANCE FOR IDENTIFICATION AND VERIFICATION TESTS ARE SHOWN AS BOLD. FOR IDENTIFICATION AND VERIFICATION EXPERIMENTS WE HAVE  $2625 \times 81$  AND  $2625 \times 81 \times 152$  TESTS RESPECTIVELY. DISTANCE BETWEEN UPPER 95% CONFIDENCE INTERVAL AND ERROR MEAN FOR IDENTIFICATION AND VERIFICATION TESTS ARE 0.1% AND 0.02% RESPECTIVELY.

		Subsystems			Score fusion		
		GMM	TVS-cosine	TVS-PLDA	mean	LR	NN
Identification Error	Overall	52.26	49.37	46.42	46.11	<b>44.25</b>	44.77
	Duration-Based	-	-	-	-	44.95	<b>44.36</b>
Verification EER	Overall	19.46	15.92	15.89	15.44	15.14	15.01
	Duration-Based	-	-	-	-	<b>13.39</b>	<b>13.46</b>

In this figure, it is observable that in most of the cells there is a small increase in the error rate. This event was expected and can be explained by significant reduction in amount of training data used for training the duration-based LR compared to overall LR. On the other hand, where the enrollment duration matches the distribution of training data (Fig. 1) and the duration of test is longer than 2 seconds, it can be seen that the duration-based LR could reduce the error by up to 66% ratio.

It also can be seen that even though by limiting the score fusion training data for each case we expect a lower accuracy gain in the duration based score fusion, the extra accuracy gained comes from special conditions where the performance of the subsystems vary a lot for different duration of enrollment and test.

## V. CONCLUSION AND FUTURE WORK

In this study we investigated the effect of duration of enrollment and test sets on different state-of-the-art speaker recognition systems. Furthermore, using several score fusion methods, we investigated the effect of score fusion of these speaker verification techniques, to benefit from the performance difference of different methods under different enrollment and test speech duration conditions. Based on our observations, duration-based technique performed better than the baseline overall score fusion methods. When we compared the accuracy of duration-based methods with overall baseline methods, it was observed that there is a dependency between the gained accuracy of TVS-based methods, and duration of enrollment and test sets, as well as the distribution of the data used for training the T matrix and gained accuracy of PLDA model are correlated. It was also observed that as the duration of enrollment and test recordings increases, the TVS-cosine and TVS-PLDA methods, with respect to GMM-UBM method give a lower gain in accuracy.

These observations motivates us to investigate the possibility of taking advantage of this performance difference of different systems for a more effective score fusion method. To this goal we investigated training a separate score fusion model for each duration condition, however this method could not give the expected gain in accuracy due to the fact that by splitting the available data for training score fusion, and training multiple models, the effectiveness of these models decrease. However, still some improvement in the overall accuracy of the systems was observed.

To prevent this condition, based on the dependence of the accuracy gain of TVS-cosine and TVS-PLDA methods to the duration of their training data, we hypothesized that there can be found a function for predicting the duration dependent weights of each subsystem

$$s = \sum_{\text{subsystems}} w_i s_i \quad (5)$$

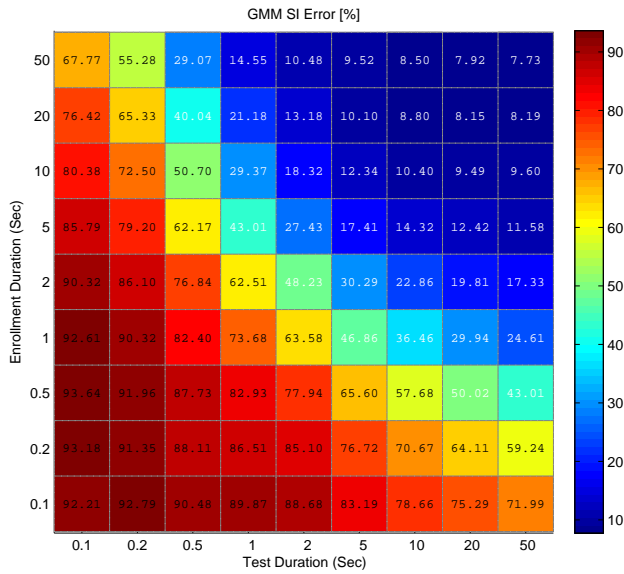
$$w_i = f_i(d_{\text{enroll}}, d_{\text{test}} | \Gamma_{\text{train}}) \quad (6)$$

where  $s$  is the final score,  $s_i$  is the output score of the  $i$ th subsystem,  $w_i$  is the corresponding weight of the  $i$ th subsystem,  $f_i$  is a subsystem specific function, mapping the duration of enrollment  $d_{\text{enroll}}$  and duration of test  $d_{\text{test}}$  and distribution of the training data  $\Gamma_{\text{train}}$  to the desired weight.

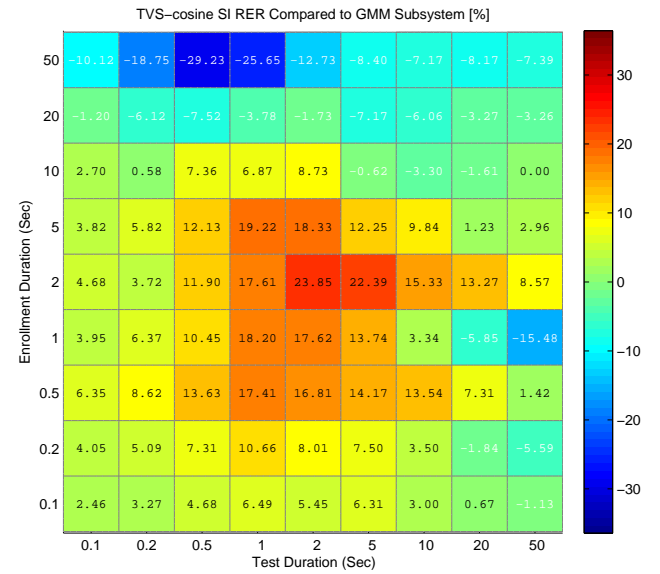
Further studies are needed to formulate such functions for each subsystem.

## REFERENCES

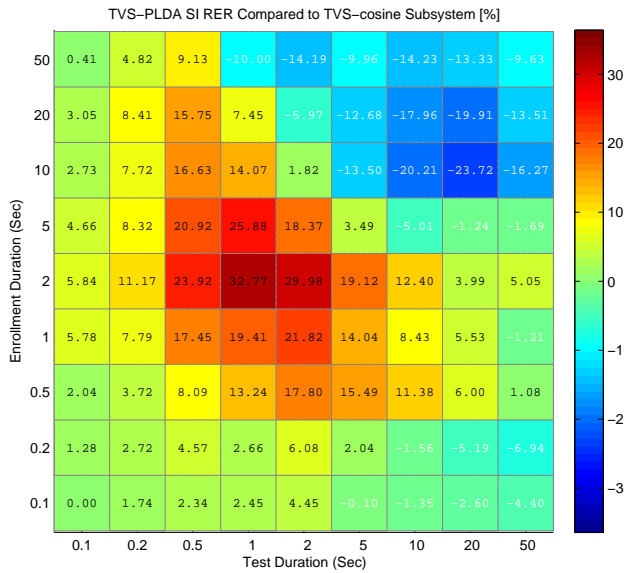
- [1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.
- [2] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," *IEEE Signal Processing Society Speech and Language Technical Committee Newsletter*, 2013.
- [3] D. Garcia-Romero and A. McCree, "Subspace-constrained supervector PLDA for speaker verification," in *INTERSPEECH*, 2013, pp. 2479–2483.
- [4] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 224–230.
- [5] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7649–7653.
- [6] S. Cumani, O. Plchot, and P. Laface, "Probabilistic linear discriminant analysis of i-vector posterior distributions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7644–7648.
- [7] B. J. Borgstrom and A. McCree, "Supervector bayesian speaker comparison," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7693–7697.
- [8] B. Vesnicer, J. Zganec-Gros, S. Dobrsek, and V. Struc, "Incorporating duration information into i-vector-based speaker-recognition systems," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 241–248.
- [9] F. Rastocanu and M. Lazar, "Score fusion methods for text-independent speaker verification applications," in *Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on*. IEEE, 2011, pp. 1–6.
- [10] V. Hautamaki, T. Kinnunen, F. Sedláč, K. A. Lee, B. Ma, and H. Li, "Sparse classifier fusion for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 1622–1631, 2013.



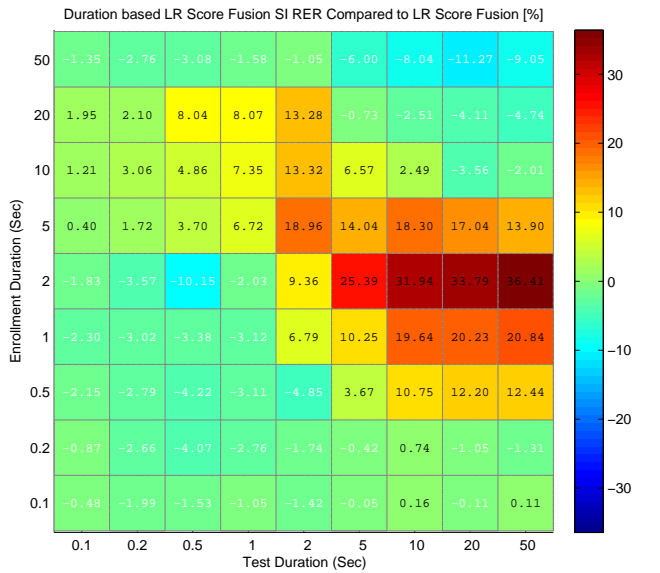
(a) GMM identification error rates for different enrollment and test duration conditions



(b) Relative error reduction of TVS-cosine compared to GMM system



(c) Relative error reduction of TVS-PLDA compared to TVS-cosine system



(d) Relative error reduction of Duration-based LR compared to LR Score Fusion

Fig. 2. Error of the GMM subsystem (a), along with the relative error reduction (RER) for speaker identification (SI) for special conditions.

- [11] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.
- [12] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, 2011, pp. 249–252.
- [13] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince, "Probabilistic models for inference about identity," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 144–157, 2012.
- [14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.
- [15] M. Westphal, "The use of cepstral means in conversational speech recognition," in *In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 1997, pp. 1143–1146.
- [16] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001, pp. 213–218.
- [17] N. Dehak and S. Shum, "Low-dimensional speech representation based on factor analysis and its applications," *Johns Hopkins CLSP Lecture*, 2011. [Online]. Available: [http://people.csail.mit.edu/sshum/talks/ivector\\_tutorial\\_interspeech\\_27Aug2011.pdf](http://people.csail.mit.edu/sshum/talks/ivector_tutorial_interspeech_27Aug2011.pdf)
- [18] D. A. Reynolds and R. C. Rose, "An integrated speech-background model for robust speaker identification," in *Acoustics, Speech and Signal Processing (ICASSP), 1992 IEEE International Conference on*, vol. 2, IEEE, 1992, pp. 185–188.
- [19] S. Goronzy and R. Kompe, "A combined MAP + MLLR approach for speaker adaptation," in *Proceedings of the Sony Research Forum*, vol. 99, 1999.
- [20] Voxforge, free speech recognition, [www.voxforge.org](http://www.voxforge.org).
- [21] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.
- [22] N. Brümmer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," *Documentation of BOSARIS toolkit*, 2011.